1 fw 2161

PTO/SB/21 (01-08)
Approved for use through 04/30/2008. OMB 0651-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE

TRANSMITTAL FORM

(to be used for all correspondence after initial filing)

Application Number	09/875,271
Filing Date	June 7, 2001
First Named Inventor	Ah H. TAN
Art Unit	2161
Examiner Name	C. Nguyen
Attorney Docket Number	455392001200
Total Number of Pages in This Submission	32

ENCLOSURES (*Check all that apply*)

- [ ] Fee Transmittal Form
  - [ ] Fee Attached
- [ ] Amendment/Reply
  - [ ] After Final
  - [ ] Affidavits/declaration(s)
- [ ] Extension of Time Request
- [ ] Express Abandonment Request
- [ ] Information Disclosure Statement
- [x] Certified Copy of Priority Document(s)
- [ ] Reply to Missing Parts/ Incomplete Application
  - [ ] Reply to Missing Parts under 37 CFR 1.52 or 1.53
- [ ] Drawing(s)
- [ ] Licensing-related Papers
- [ ] Petition
- [ ] Petition to Convert to a Provisional Application
- [ ] Power of Attorney, Revocation Change of Correspondence Address
- [ ] Terminal Disclaimer
- [ ] Request for Refund
- [ ] CD, Number of CD(s) ______
  - [ ] Landscape Table on CD
- [x] After Allowance Communication to TC
- [ ] Appeal Communication to Board of Appeals and Interferences
- [ ] Appeal Communication to TC **(Appeal Notice, Brief, Reply Brief)**
- [ ] Proprietary Information
- [ ] Status Letter
- [ ] Other Enclosure(s) (please Identify below):

Remarks

SIGNATURE OF APPLICANT, ATTORNEY, OR AGENT

Firm Name	MORRISON & FOERSTER LLP		
Signature	Deborah S Gladstein		
Printed name	Deborah S. Gladstein		
Date	May 5, 2008	Reg. No.	43,636

Docket No.: 455392001200
(PATENT)

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:
Ah H. TAN

Application No.: 09/875,271

Confirmation No.: 4593

Filed: June 7, 2001

Art Unit: 2161

For: METHOD AND SYSTEM FOR USER-CONFIGURABLE CLUSTERING OF INFORMATION

Examiner: C. Nguyen

CLAIM FOR PRIORITY AND SUBMISSION OF DOCUMENTS

MS Issue Fee
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Sir:

Applicant hereby claims priority under 35 U.S.C. 119 based on the following prior foreign application filed in the following foreign country on the date indicated:

Country	Application No.	Date
Singapore	200003177-3	June 7, 2000

In support of this claim, a certified copy of the said original foreign application is filed herewith.

Dated: May 5, 2008

Respectfully submitted,

By Deborah S. Gladstein

Deborah S. Gladstein
Registration No.: 43,636
MORRISON & FOERSTER LLP
1650 Tysons Blvd, Suite 400
McLean, Virginia 22102
(703) 760-7753

THE REGISTRY OF PATENTS
SINGAPORE

THE PATENTS ACT
(CHAPTER 221)

CERTIFICATE OF GRANT OF PATENT

In accordance with section 35 of the Patents Act, it is hereby certified that a patent having the P-No. 93868 has been granted in respect of an invention having the following particulars:

Title	:	METHOD AND SYSTEM FOR USER-CONFIGURABLE CLUSTERING OF INFORMATION
Application Number	:	200003177-3
Date of Filing	:	07 June 2000
Priority Data	:	-
Name of Inventor(s)	:	TAN, AH HWEE
Name(s) and Address(es) of Proprietor(s) of Patent	:	KENT RIDGE DIGITAL LABS 21 HENG MUI KENG TERRACE SINGAPORE 119613
Date of Grant	:	30 December 2004

Dated this 30th day of December 2004.

Liew Woon Yin (Ms)
Registrar of Patents,
Singapore.

07 JUN 2000

METHOD AND SYSTEM FOR USER-CONFIGURABLE CLUSTERING OF INFORMATION

FIELD OF THE INVENTION

This invention relates to the field of pattern processing and information management and more specifically to methods and systems for incorporating user interests and preferences in automated information clustering. Related fields of invention include information database content management, self-organizing information databases, document clustering, and personalized systems.

BACKGROUND OF THE INVENTION

Categorization and clustering have been two fundamental approaches to information organization and information database content management.

Categorization or classification is supervised in nature. A user defines a fixed number of classes or categories. The task is to assign a pattern or object to one or more of the classes. Categorization provides good control in the sense that it organizes the information according to the structure defined by the user. However, due to the predefined structure, categorization is not well suited to handling novel data. In addition, much effort is needed to build a categorization system. It is necessary to specify classification knowledge in terms of classification rules or keywords (disclosed in U.S. Pat. No. 5,371,807) or to construct a categorization system through some supervised learning algorithms (disclosed in U.S. Pat. Nos. 5,671,333 and 5,675,710). The former requires knowledge specification (*e.g.*, written classification rules) and the latter requires example annotations (*i.e.* labeling information). Both are labor intensive.

Clustering is unsupervised in nature. For unsupervised systems (U.S. Pat. Nos. 5,857,179 and 5,787,420), there is no need to train or construct a classifier since information is organized automatically into groups based on similarities. However, a user has very little control over how the information is grouped together. Although it is possible to fine tune the parameters of the similarity measures to control the degree of coarseness, the effect of changing a parameter cannot be predicted; changing one

parameter could affect all clustered results. In addition, the structure established through the clustering process is unpredictable. Whereas clustering is acceptable for a pool of relatively static information, in situations where new information is received every day, information with similar content may be grouped (based on different themes) into different clusters on different days. This ever-changing cluster structure is highly undesirable for the user who is navigating the information database to find desired information. Imagine the frustration of reading a newspaper with a different layout every day! U.S. Pat. No. 5,911,140 attempts to provide a solution by ordering document clusters based on user interests. However, the cluster ranking relies on the availability of the ranking of each document in the clusters and only very minimal user preferences are taken into account.

SUMMARY OF THE INVENTION

In order to overcome the various shortcomings of systems which effect only categorization or clustering of information with little or no account taken of user preferences, this invention provides a method and system that incorporate users' preferences in an information clustering system. In general, this system allows a user to create a cluster structure and influence or personalize the cluster structure by indicating his or her own preferences as to how information should be grouped. This invention further allows the user to store the cluster structure and subsequently retrieve it for future use.

The user-configurable information clustering system comprises an *information clustering engine* for clustering of information based on similarities, a *user interface module* for displaying the information groupings and obtaining user preferences, a *personalization module* for defining, labeling, modifying, storing and retrieving cluster structure, and a *knowledge base* where a user-defined cluster structure is stored.

According to the invention, each unit of information is represented by an *information vector*. A user preference, indicating a preferred grouping for the corresponding unit of information, can be represented by a *preference vector*. In addition, information, which may be in the form of a database, is supplied to the user-configurable information clustering system by any well known means within the art.

In the preferred embodiment, the information clustering engine is a hybrid neural network comprising two input fields F_1^a and F_1^b with an F_2 cluster field. The F_1^a field serves as the input field for the information vector **A**. The F_1^b field serves as the input field for the preference vector **B**. The F_2 field contains a plurality of cluster nodes, each encoding a template information vector w_j^a and a template preference vector w_j^b. Given an information vector **A** with an associated preference vector **B**, the system first searches for an F_2 cluster J encoding a template information vector w_J^a that is closest to the information vector **A** according to a similarity function. It then checks if the associated F_2 template preference vector w_J^b of the selected category matches with the input preference vector. If so, the templates of the F_2 cluster J are modified to encode the input information and preference vectors. Otherwise, the cluster is reset and the system repeats the process until a match is found.

Through the user interface and the personalization module, the user is able to influence the cluster structure by indicating his own preferences in the form of preference vectors. The user can create a new cluster, label an existing cluster, and/or modify cluster structure by merging and splitting clusters. In addition, the resulting customized cluster structure can be stored in the *cluster structure knowledge base* and retrieved at a later stage for processing new information.

In one embodiment of the invention, a system and method are provided for customizing the organization of an existing set of information according to the user's knowledge and preferences. In another embodiment, a system and method are provided for creating a cluster structure automatically and subsequently modifying this machine-generated cluster structure according to the user's preferences. In yet another embodiment, a system and method are provided for detecting new information and analyzing trends wherein the user, through repeated personalization of the cluster structure, identifies new information and trends previously unknown to him or her.

The disclosed invention is more flexible than a pure categorization system in which information must be assigned to one or more pre-defined categories or groups. At the same time, it is more flexible than a pure clustering or self-organizing system in which information is grouped according to similarities but the user has very little control over how the information is organized.

The invention has a number of advantages over the prior art: The invention performs clustering or self-organization of information based on similarities in content, *i.e.* similarities in information vectors. The information can be automatically organized without user training or prior construction of a classifier. The invention allows the user to correct or change the organization of the information as necessary. The invention also allows the user to intervene in the organization of the information both globally and locally. Further, the invention allows the user to control the coarseness of the information groupings without tuning the parameters of complex similarity functions. The invention also allows the user to indicate directly how specific units of information are to be organized.

BRIEF DESCRIPTION OF THE DRAWINGS

Embodiments of the invention will now be described by way of example with reference to the accompanying drawings in which:

FIG. 1 illustrates an embodiment of a user-configurable information clustering system according to the present invention.

FIG. 2A shows an exemplary architecture for the information clustering engine of FIG. 1 for clustering information in response to user preferences. Information is encoded as information vectors. User preferences are encoded as preference vectors. The F_1^a field serves as the input field for the information vectors. The F_1^b field serves as the input field for the preference vectors. Information clusters (represented as F_2) are formed through the synchronized clustering of the information and preference vectors.

FIG. 2B illustrates the category choice process of the information clustering engine of FIG. 1.

FIG. 2C illustrates the template matching and the template learning process of the information clustering engine of FIG. 1.

FIG. 3 illustrates an exemplary flow diagram for incorporating user preferences in the information clustering process.

FIG. 4 illustrates an exemplary set of personalization functions that a user could use while performing information clustering.

DETAILED DESCRIPTION OF THE INVENTION

The invention is concerned with the organization of information, based on similarities in content, *i.e.* similarities in information vectors, according to a user's knowledge and preferences. The information comprises text, image, audio, video, or any combinations thereof. According to the invention, each unit of information (Φ), defined as an individual element of information, may be any object, for example a document, person, company, country, etc., and can be represented by a complement coded 2M-dimensional information vector A of *attributes* or *features*,

$$\mathbf{A} = (\mathbf{a}, \mathbf{a}^c) = (a_1,\ldots,a_M,a_1^c,\ldots,a_M^c) \tag{1}$$

where a_i is a real-valued number between zero and one, indicating the degree of presence of attribute i, and $a_i^c = 1 - a_i$. Complement coding represents both the on-response and the off-response to an input vector and preserves amplitude information upon normalization.

In the case of documents, for example, the features in the information vectors could be token words commonly known as *keywords*. For example, in the exemplary case where the information is a business document, keywords may include "share", "market", "stock", "acquisition", "trading" etc. The feature sets can be predefined manually or generated automatically from the information set.

User preferences are represented by *preference vectors* that indicate the preferred groupings of the information. A preference vector B is a complement coded 2N-dimensional vector defined as

$$\mathbf{B} = (\mathbf{b}, \mathbf{b}^c) = (b_1,\ldots,b_N,b_1^c,\ldots,b_N^c) \tag{2}$$

where b_k is either zero or one, indicating the presence or absence of a user-defined class label L_k, and $b_k^c = 1 - b_k$. A user's knowledge comprises that information acquired by one in the user's profession, community, or field of endeavor or study. Also, this knowledge may include highly specialized information developed, acquired or used based on a user's unique experiences. An example of user knowledge includes a physician's knowledge of diseases, symptoms for those diseases, and appropriate treatments, including appropriate drugs; knowledge of this type from many practitioners could be placed in a database which could increase in content as new diseases were discovered, treated and cured.

Using the invention, a physician could configure the database and additions thereto over time according to her own experience and specialty. Another example of user knowledge is a market analyst's understanding of trends in a given manufacturing sector over a given period of time; a database of wire news reports over a given period or updated on a periodic basis may contain information of use to the market analyst. Using the invention, the market analyst could configure such a database and updates thereto according to the particular sector of interest to him. Yet another example of user knowledge is a chef's knowledge of the cooking arts; a database of culinary dishes, ingredients, foods, etc., which may be intermittently modified, contains information of use to the culinary field. Using the invention, the chef can, for example, create groupings of ingredients for culinary dishes of interest to her. Another example of user knowledge is the preferred grouping of news articles by a journalist. A journalist, depending on his target readers, may organize articles from local and foreign sources into specific groupings that would be most appropriate for his purposes. For example, a Singapore journalist might organize foreign news into threads that are of interest to Singaporeans, such as "Michael Fay Event", "Singapore Economics" etc. Other examples will be evident to those skilled in the art. In addition, the user's preferences are derived from, for example, his personal or professional informational desires, organizational goals and objectives, analytical training and skills, interpretational biases, and experiences with identification of information. User preferences include informational and organizational objectives useful to organizing and increasing the knowledge base of the user. For example, in the case of the Singapore journalist mentioned above, preferences used in organizing the articles are based on the journalist's interpretation and experience on the relevancy of such topics to his readers and the volume of articles on such topics over time.

Units of information which share common attributes or content, as indicated by their information vectors, are said to be similar to each other. Units of information which share few or no common attributes or content are said to be dissimilar to each other. The system may automatically determine similarity between two units of information based on the application of a similarity function. Alternately, the user may determine the degree to which units of information appear to be similar to each other based on his or her knowledge and preferences.

The disclosed method can be executed using a computer system, such as a personal computer or the like, as is well known in the art. The disclosed system can be a stand-alone system, or it can be incorporated in a computer system, in which case the user interface can be the graphical or other user interface of the computer system and the cluster structure knowledge base can be, for example, a file in any of the computer system's storage areas, elements or devices.

Referring to **FIG. 1**, there is provided a user-configurable information clustering system **10** comprising an *information clustering engine* **20**, a *personalization module* **50**, a *cluster structure knowledge base* **14**, and a *user interface module* **16**. Information in the form of information and preference vectors is supplied to the clustering engine **20**, which comprises a hybrid neural network model that performs a hybrid of supervised and unsupervised learning. The neural network may be conventional. For example, an ARTMAP system or an ARAM system, such as described in "Adaptive Resonance Associative Map", published in "Neural Networks", Vol. 8 No. 3, pp. 437-446 (1995), which is incorporated herein by reference, can be used. The user interface module **16** may comprise a graphical user interface, keyboard, keypad, mouse, voice command recognition system, or any combination thereof, and may permit graphical visualization of information groupings. The cluster structure knowledge base **14** may be any conventional recordable storage format, for example a file in a storage device, such as magnetic or optical storage media, or in a storage area of a computer system.

The user-configurable information clustering system **10** allows a user to personalize or influence the cluster structure by indicating his or her own preferences in the form of preference vectors. Through the *user interface module* **16** and the *personalization module* **50**, the user is able to create a customized cluster structure by selective and/or repeated application of the following: creating a new cluster, labeling an existing cluster, and modifying cluster structure by merging and splitting clusters. In addition, the customized cluster structure can be stored in the *cluster structure knowledge base* **14** and retrieved at a later stage for processing new information.

As described in the article cited above, ARAM is a family of neural network models that performs incremental supervised learning of recognition categories (pattern classes) and multidimensional maps of both binary and analog patterns. Referring to **FIG.**

2A, an ARAM system can be visualized as two overlapping Adaptive Resonance Theory (ART) modules consisting of two input fields F_1^a **22** and F_1^b **26** with a cluster field F_2 **30**.

Each F_2 cluster node j is associated with an adaptive template information vector w_j^a and corresponding adaptive template preference vector w_j^b. Initially, all cluster nodes are uncommitted and all weights are set equal to 1. After a cluster node is selected for encoding, it becomes committed.

Fuzzy ARAM dynamics are determined by the choice parameters $\alpha^a>0$ and $\alpha^b>0$; the learning rates β^a in [0,1] and β^b in [0,1]; the vigilance parameters ρ^a **24** in [0,1] and ρ^b **28** in [0,1]; and a contribution parameter γ in [0,1]. The choice parameters α^a and α^b control the bias towards choosing a F_2 cluster whose template information and preference vectors have a larger norm or magnitude. The learning rates β^a and β^b control how fast the template information and preference vectors w_j^a and w_j^b adapt to the input information and preference vectors **A** and **B**, respectively. The vigilance parameters ρ^a and ρ^b determine the criteria for a satisfactory match between the input and the template information and preference vectors, respectively. The contribution parameter γ controls the weighting of contribution from the information and preference vectors when selecting an F_2 cluster.

Referring to **FIG. 2B**, given an information vector **A** with an associated preference vector **B**, the system first searches for an F_2 cluster J encoding a template information vector w_J^a and a template preference vector w_J^b paired therewith that are closest to the input information vector **A** and the input preference vector **B**, respectively, according to a similarity function. Specifically, for each F_2 cluster j, the information clustering engine calculates a similarity score based on the input information and preference vectors **A** and **B**, respectively, and the template information and preference vectors w_j^a and w_j^b, respectively. An example of a similarity function is given below as the *category choice* function, eqn. (3). The F_2 cluster that has the maximal similarity score is then selected and indexed at J.

Referring to **FIG. 2C**, the information clustering engine performs *template matching* to verify that the template information vector w_J^a and the template preference vector w_J^b of the selected category J match well with the input information vector **A** and the input preference vector **B**, respectively, according to another similarity function, *e.g.*

eqn. (4) below. If so, the system performs *template learning* to modify the template vectors $\mathbf{w}_J^a$ and $\mathbf{w}_J^b$ of the F_2 cluster J to encode the input information and preference vectors **A** and **B**, respectively. Otherwise, the cluster is reset and the system repeats the process until a match is found. The detailed algorithm is given below.

The ART modules used in ARAM may be of a type which categorizes binary patterns, analog patterns, or a combination of the two patterns (referred to as "fuzzy ART"), as is known in the art. Described below is a fuzzy ARAM model composed of two overlapping fuzzy ART modules.

Referring to **FIG. 3**, the dynamics of the information clustering engine **20** is described as follows. Given a pair of F_1^a and F_1^b input vectors **A** and **B**, for example, an information vector and preference vector, respectively, for each F_2 node j, a *category choice process* **32** computes the choice function T_j as defined by

$$T_j = \gamma\, |\mathbf{A} \wedge \mathbf{w}_j^a| \,/\, (\alpha^a + |\mathbf{w}_j^a|) + (1-\gamma)\, |\mathbf{B} \wedge \mathbf{w}_j^b| \,/\, (\alpha^b + |\mathbf{w}_j^b|) \qquad (3)$$

where, for vectors **p** and **q**, the fuzzy AND operation is defined by $(\mathbf{p} \wedge \mathbf{q})_i = \min(p_i, q_i)$, and the norm is defined by $|\mathbf{p}| = \Sigma_i\, p_i$. The system is said to make a choice when at most one F_2 node can become active. The choice is indexed at J by a *select winner process* **34** where $T_J = \max\{T_j$: for all F_2 nodes j$\}$.

A *template matching process* **36** then checks if the selected cluster represents a good match. Specifically, a check **38** is performed to verify if the *match functions*, m_J^a and m_J^b, meet the vigilance criteria in their respective modules:

$$m_J^a = |\mathbf{A} \wedge \mathbf{w}_J^a| \,/\, |\mathbf{A}| \geq \rho^a \text{ and } m_J^b = |\mathbf{B} \wedge \mathbf{w}_J^b| \,/\, |\mathbf{B}| \geq \rho^b. \qquad (4)$$

Resonance occurs if both criteria are satisfied. Learning then ensues, as defined below. If any of the vigilance constraints is violated, mismatch reset **42** occurs in which the value of the choice function T_J is set to 0 for the duration of the input presentation. The search process repeats, selecting a new index J until resonance is achieved.

Once the search ends, a *template learning process* **40** updates the template information and preference vectors $\mathbf{w}_J^a$ and $\mathbf{w}_J^b$, respectively, according to the equations

$$\mathbf{w}_J^{a\,(new)} = (1-\beta^a)\, \mathbf{w}_J^{a\,(old)} + \beta^a\, (\mathbf{A} \wedge \mathbf{w}_J^{a\,(old)}) \qquad (5)$$

and

$$w_J^{b\,(new)} = (1-\beta^b)\, w_J^{b\,(old)} + \beta^b\ (\mathbf{B} \wedge w_J^{b\,(old)}) \tag{6}$$

respectively. For efficient coding of noisy input sets, it is useful to set $\beta^a = \beta^b = 1$ when J is an uncommitted node, and then take $\beta^a < 1$ and $\beta^b < 1$ after the cluster node is committed. *Fast learning* corresponds to setting $\beta^a = \beta^b = 1$ for committed nodes.

At the start of each input presentation, the vigilance parameter ρ_a equals a baseline vigilance $\bar{\rho}^a$. If a reset occurs in the category field F_2, a *match tracking process* **44** increases ρ^a until it is slightly larger than the match function m_J^a. The search process then selects another F_2 node J under the revised vigilance criterion.

Referring to **FIG. 4**, the *personalization module* **50** works in conjunction with the *information clustering engine* **20** to incorporate user preferences to modify the machine-generated cluster structure.

An exemplary parameter setting for the information clustering engine **20** is as follows: $\alpha^a=\alpha^b=0.1$, $\beta^a=\beta^b=1$, $\bar{\rho}^a=0.5$, $\rho^b=1$, and $\gamma=0.5$. During automatic clustering, no user preference is given, and the information clustering engine **20** automatically generates a cluster structure, referred to as a machine-generated cluster structure: For each unit of information (Φ), a pair of vectors ($\mathbf{A}$, $\mathbf{B}_0$) is presented to the system, where $\mathbf{A}$ is the representation vector of Φ and

$$B_{0i} = 1 \text{ for } i = 1,\ldots,2N. \tag{7}$$

Since $|\mathbf{B}_0 \wedge w_j^b| / |\mathbf{B}_0|$ equals 1, condition (6) reduces to

$$m_j^a = |\mathbf{A} \wedge w_j^a| / |\mathbf{A}| \geq \rho^a. \tag{8}$$

Essentially, the system now operates like a pure clustering system that self-organizes the information based on similarities in the information vectors. The coarseness of the information groupings is controlled by the baseline vigilance parameter ($\bar{\rho}^a$).

A *create cluster module* **52** allows the user to add a new information cluster into the system so that information can be organized according to such an information grouping. Through the *user interface module* **16**, the user can input a pair of template information and preference vectors (w_J^a, w_J^b) which defines the key attributes of the information in the cluster together with a cluster label, if any. The resulting clusters

reflect the user's preferred way of grouping information and can be used as the default slots for organizing information.

A *label cluster module* **54** allows the user to assign labels to "mark" certain information groupings that are of particular interest (to the user) so that new information can be organized according to such information groupings. Through the *user interface module* **16**, the user can assign a label L_k to a cluster j by modifying the template preference vector $\mathbf{w}_j^b$ to equal $\mathbf{B}_k$, where $\mathbf{B}_k$ is a preference vector representing L_k. Labels reflect the user's interpretation of the groupings. They are useful landmarks to the user in navigating the information database and locating old as well as new information.

Using the *label cluster module* **54**, the user is able to merge clusters implicitly by labeling them with the same labels. In this case, the merging is said to be a local one as it only affects the clusters that are labeled. To do a global merging, a *merge cluster module* **56** allows the user to combine two or more information groupings generated by the clustering process using a lower vigilance parameter value. Through the *user interface module* **16**, the user can select one or more units of information in each of two different clusters as an indicative standard of similarity. As an example, $\mathbf{A}_1$ and $\mathbf{A}_2$ can be two information vectors representing two units of information. The revised baseline vigilance parameter ρ^a would then be computed as

$$\rho^a = \min\left(|\mathbf{A}_1 \wedge \mathbf{A}_2| / |\mathbf{A}_1|, |\mathbf{A}_1 \wedge \mathbf{A}_2| / |\mathbf{A}_2|\right). \tag{9}$$

Using the new baseline vigilance parameter, $\mathbf{A}_1$ and $\mathbf{A}_2$ will satisfy the match condition as stated in (4) and may be grouped into one cluster as a result of the relaxed similarity criteria. The effect of cluster merging is global, in the sense that the system now operates at a lower vigilance on the whole, grouping items together that it would otherwise distinguish.

A *split cluster module* **58** allows the user to split an information group into two or more clusters by indicating that certain units of information are sufficiently different to be grouped separately. Through the *user interface module* **16**, the user can select two specific units of information, for example $\mathbf{A}_1$ and $\mathbf{A}_2$, in a cluster and assign to them two different labels, for example L_1 and L_2, represented by $\mathbf{B}_1$ and $\mathbf{B}_2$, respectively.

The updated pairs of information and preference vectors, ($\mathbf{A}_1$, $\mathbf{B}_1$) and ($\mathbf{A}_2$, $\mathbf{B}_2$), together with the rest of the vectors, are presented to the information clustering engine **20** for re-clustering. Since $\mathbf{B}_1 \neq \mathbf{B}_2$, and $\rho^b=1$, $\mathbf{A}_1$ and $\mathbf{A}_2$ will be grouped into different clusters. In addition, the remaining information vectors, originally in a single cluster, will be re-organized into one of the two new clusters based on their similarities to $\mathbf{A}_1$ and $\mathbf{A}_2$.

In another example, $\mathbf{A}_1$ and $\mathbf{A}_2$ can be used to tighten the match condition for the entire information space. In this case, the baseline vigilance parameter ρ^a would be re-computed as

$$\rho^a = \max\left(|\mathbf{A}_1 \wedge \mathbf{A}_2| / |\mathbf{A}_1|, |\mathbf{A}_1 \wedge \mathbf{A}_2| / |\mathbf{A}_2|\right) + \varepsilon, \tag{12}$$

where ε is a small constant.

By selective and/or repeated application of the create cluster, label cluster, merge cluster and split cluster functions of the personalization module **50**, the user is able to create a labeled cluster structure which incorporates his own preferences.

To preserve the labeled cluster structure for a fresh clustering session, a *store cluster module* **60** transfers the following information to the cluster structure knowledge base **14**.

1. The baseline vigilance parameter (ρ^a).
2. For each labeled cluster j, the template information vector $\mathbf{w}_j^a$ and the associated template preference vector $\mathbf{w}_j^b$.

The stored clusters and cluster structure can be retrieved at a later stage from the knowledge base **14** using a *retrieve cluster module* **62** to initialize the user-configurable information clustering system, that is, the retrieved cluster structure is used as the initial architecture or structure of the information clustering engine **20**. After initialization of the information clustering system based on the retrieved cluster structure, new information can be organized according to the user's preferences stored over the previous sessions. The cluster structure may also be modified by further personalization.

An embodiment of the disclosed invention is *personalized document navigation* wherein the user is allowed to customize the document navigation space, *i.e.* the organized collection of information available to the user, with respect to his or her interpretation and preferences.

Another embodiment of the invention is a *drag-and-draw approach to building a categorization system*. Information is first automatically clustered into natural groupings based on similarities in content. By modifying the machine-generated groupings, the user can define her preferred groupings by selective and/or repeated application of the personalization functions described herein. A classification system can be created in an intuitive and interactive manner without the need for example annotation (i.e., labeling information) or knowledge specification (e.g., written classification rules).

Yet another embodiment is *detection of new information and trend analysis*. The user defines his know-how and interpretation of the environment in terms of how he wants the information to be organized. New information is supplied periodically to the information clustering system. New information that falls within the user-defined cluster structure corresponds to familiar themes of information. Any new information that falls outside of the defined cluster structure represents new themes which are potentially interesting to the user. Repeated personalization of the information in the information database, i.e. selectively and/or repeatedly creating, labeling, merging, splitting, and storing clusters and the resulting labeled cluster structure, helps the user to identify information that is new with respect to his experience and to analyze unexpected trends.

Various preferred embodiments of the invention have now been described. While these embodiments have been set forth by way of example, various other embodiments and modifications will be apparent to those skilled in the art. Accordingly, it should be understood that the invention is not limited to such embodiments, but encompasses all that which is described in the following claims.

What is claimed is

1. A method of organizing information into a plurality of classes or clusters with a user-configurable information clustering system comprising:
   a) grouping units of information into clusters based on similarities to create a cluster structure; and
   b) personalizing said cluster structure according to user knowledge and preferences.

2. The method according to claim 1 wherein said grouping units of information into clusters is carried out automatically to create a machine-generated cluster structure.

3. The method according to claim 1 wherein said personalizing comprises creating at least one new information cluster.

4. The method according to claim 3 wherein said personalizing further comprises labeling each information cluster.

5. The method according to claim 4 wherein said personalizing further comprises merging information clusters.

6. The method according to claim 5 wherein said personalizing further comprises splitting at least one information cluster.

7. The method according to claim 6 wherein said personalizing further comprises storing said cluster structure in a knowledge base.

8. The method according to claim 7 wherein said personalizing further comprises retrieving said cluster structure from said knowledge base to initialize said user-configurable information clustering system prior to clustering new information.

9. The method according to claim 1 wherein said personalizing comprises labeling each information cluster.

10. The method according to claim 1 wherein said personalizing comprises merging information clusters.

11. The method according to claim 1 wherein said personalizing comprises splitting at least one information cluster.

12. The method according to claim 1 wherein said personalizing comprises storing said cluster structure in a knowledge base.

13. The method according to claim 1 wherein said personalizing comprises retrieving said cluster structure from a knowledge base to initialize said user-configurable information clustering system prior to clustering new information.

14. The method according to claim 1 wherein said information comprises text, image, audio, video or any combination thereof.

15. The method according to claim 1 wherein said user-configurable information clustering system comprises an adaptive resonance associative map.

16. The method according to claim 1 wherein said user-configurable information clustering system incorporates user knowledge and preferences for information clustering.

17. The method according to claim 1 wherein said user-configurable information clustering system further comprises a user interface.

18. The method according to claim 1 wherein each of said units of information is represented by an information vector.

19. The method according to claim 1 wherein a user-preferred information grouping is represented by a preference vector.

20. The method according to claim 1 wherein said units of information are grouped into classes or clusters based on a similarity function.

21. The method according to claim 20 wherein said classes or clusters have a coarseness which is controlled by a baseline vigilance parameter.

22. The method according to claim 1 further comprising indication by a user of a preference for a lower baseline vigilance parameter by selecting at least one unit of information from each of at least two clusters wherein the selected units of information are deemed by the user to be similar to each other.

23. The method according to claim 1 further comprising indication by a user of a preference for a higher baseline vigilance parameter by selecting at least two units of information in a cluster, wherein said units of information are deemed by the user to be dissimilar to each other.

24. A method of building an information classification system comprising:
   a) grouping units of information into clusters based on similarities to create a cluster structure;
   b) personalizing said information clusters according to a user's knowledge and preferences; and
   c) storing said personalized clusters as the defined categories of a classification system.

25. The method according to claim 24 wherein said personalizing comprises labeling each information cluster.

26. The method according to claim 24 wherein said personalizing comprises creating at least one new information cluster.

27. The method according to claim 24 wherein said personalizing comprises merging information clusters.

28. The method according to claim 24 wherein said personalizing comprises splitting at least one information cluster.

29. The method according to claim 24 wherein said information comprises text, image, audio, video or any combination thereof.

30. A method for new information detection and trend analysis with a user-configurable information clustering system comprising:
    a) grouping units of information into clusters based on similarities to create a cluster structure;
    b) personalizing said cluster structure according to a user's knowledge and preferences;
    c) storing said personalized cluster structure, wherein said personalized cluster structure defines the user's knowledge of said units of information;
    d) retrieving said stored cluster structure to initialize said information clustering system;
    e) clustering new information, using said initialized information clustering system; and
    f) analyzing new clusters and said retrieved cluster structure, wherein said new clusters are grouped according to the user's preferences.

31. The method according to claim 30 wherein said personalizing comprises labeling each information cluster.

32. The method according to claim 30 wherein said personalizing further comprises creating at least one new information cluster.

33. The method according to claim 30 wherein said personalizing further comprises merging information clusters.

34. The method according to claim 30 wherein said personalizing further comprises splitting at least one information cluster.

35. The method according to claim 30 wherein said information comprises text, image, audio, video or any combination thereof.

36. A user-configurable information clustering system comprising:

a) an information clustering engine for clustering units of information based on similarities to create a cluster structure;

b) a personalization module for personalizing said cluster structure according to user knowledge and preferences;

c) a user interface; and

d) a knowledge base for storing said cluster structure.

37. The system according to claim 36 wherein said information clustering engine automatically clusters information to create a machine-generated cluster structure.

38. The system according to claim 36 wherein said personalization module comprises means for creating at least one new information cluster.

39. The system according to claim 38 wherein said personalization module further comprises means for labeling each information cluster.

40. The system according to claim 39 wherein said personalization module further comprises means for merging information clusters.

41. The system according to claim 40 wherein said personalization module further comprises means for splitting at least one information cluster.

42. The system according to claim 41 wherein said personalization module further comprises means for storing the cluster structure in said knowledge base.

43. The system according to claim 42 wherein said personalization module further comprises means for retrieving the cluster structure from said knowledge base to initialize said user-configurable information clustering system prior to clustering new information.

44. The system according to claim 36 wherein said personalization module comprises means for labeling each information cluster.

45. The system according to claim 36 wherein said personalization module comprises means for merging information clusters.

46. The system according to claim 36 wherein said personalization module comprises means for splitting at least one information cluster.

47. The system according to claim 36 wherein said personalization module comprises means for storing the cluster structure in said knowledge base.

48. The system according to claim 36 wherein said personalization module comprises means for retrieving the cluster structure from said knowledge base to initialize said user-configurable information clustering system prior to clustering new information.

49. The system according to claim 36 wherein said information comprises text, image, audio, video or any combination thereof.

50. The system according to claim 36 wherein user knowledge and preferences are incorporated in information clustering.

51. The system according to claim 36 wherein said information clustering engine comprises an adaptive resonance associative map.

52. The system according to claim 36 wherein said user interface permits graphical visualization of said information clusters.

53. The system according to claim 36 wherein each of said units of information is represented by an information vector.

54. The system according to claim 36 wherein a user-preferred information grouping is represented by a preference vector.

55. The system according to claim 36 wherein said units of information are grouped into classes or clusters based on a similarity function.

56. The system according to claim 55 wherein said classes or clusters have a coarseness which is controlled by a baseline vigilance parameter.

57. The system according to claim 36 wherein said personalization module permits indication by a user of a preference for a lower baseline vigilance parameter by selecting at least one unit of information from each of at least two clusters wherein said selected units of information are deemed by the user to be similar to each other.

58. The system according to claim 36 wherein said personalization module permits indication by a user of a preference for a higher baseline vigilance parameter by selecting at least two units of information in a cluster, wherein said units of information are deemed by the user to be dissimilar to each other.

59. An information classification system comprising:
   a) means for grouping information vectors into clusters based on similarities to create a cluster structure;
   b) means for personalizing said information clusters according to a user's knowledge and preferences; and
   c) means for storing said personalized clusters as the defined categories of a classification system.

60. The system according to claim 59 wherein said personalizing means permits labeling each information cluster.

61. The system according to claim 59 wherein said personalizing means permits creating a plurality of new information clusters.

62. The system according to claim 59 wherein said personalizing means permits merging information clusters.

63. The system according to claim 59 wherein said personalizing means permits splitting at least one information cluster.

64. The system according to claim 59 wherein the information comprises text, image, audio, video or any combination thereof.

65. A system for new information detection and trend analysis comprising:
   a) means for grouping units of information into clusters based on similarities;
   b) means for personalizing said clusters according to a user's knowledge and preferences;
   c) means for storing said personalized clusters wherein said personalized clusters define a user's knowledge of said units of information; and

d) means for retrieving said personalized clusters prior to clustering new information to permit analysis of new clusters and said retrieved clusters, wherein said new clusters are grouped according to the user's preferences.

66. The system according to claim 65 wherein said personalizing means permits labeling each information cluster.

67. The system according to claim 65 wherein said personalizing means permits creating a plurality of information clusters.

68. The system according to claim 65 wherein said personalizing means permits merging information clusters.

69. The system according to claim 65 wherein said personalizing means permits splitting at least one information cluster.

70. The system according to claim 65 wherein the information comprises text, image, audio, video or any combination thereof.

DRAWINGS

FIG. 1

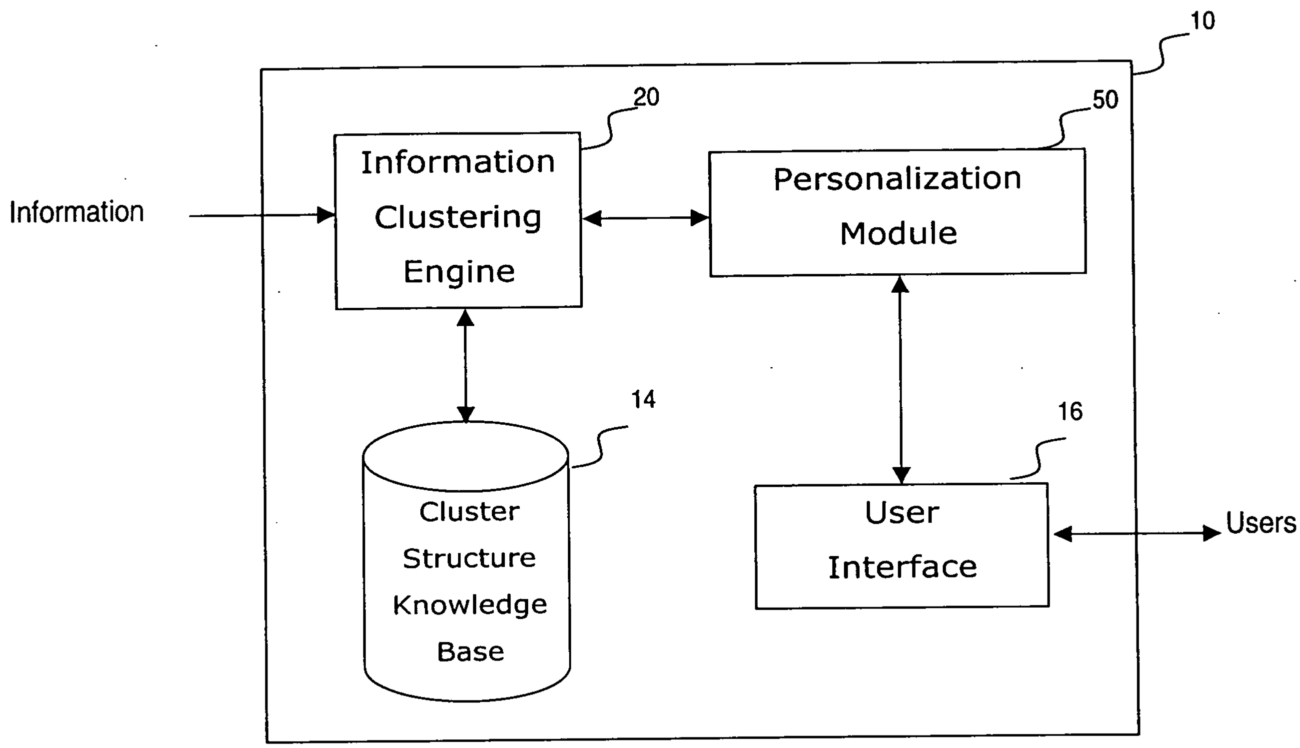
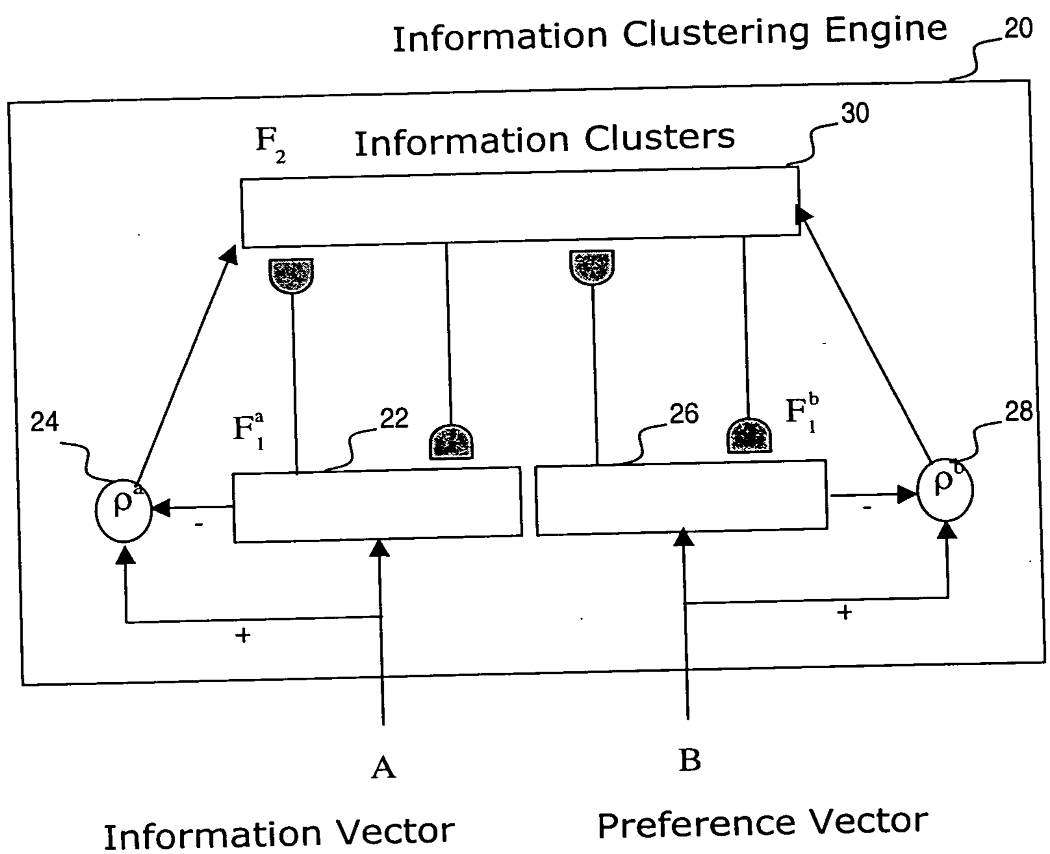


FIG. 2A



Information Clustering Engine
20
F2
Information Clusters
30
24
ρa
Fa1
22
26
Fb1
28
ρb
-
-
+
+
A
B
Information Vector
Preference Vector

FIG. 2B

Category choice $$T_j = \gamma \frac{|A \wedge W_j^a|}{\alpha^a + |W_j^a|} + (1-\gamma)\frac{|B \wedge W_j^b|}{\alpha^b + |W_j^b|}$$

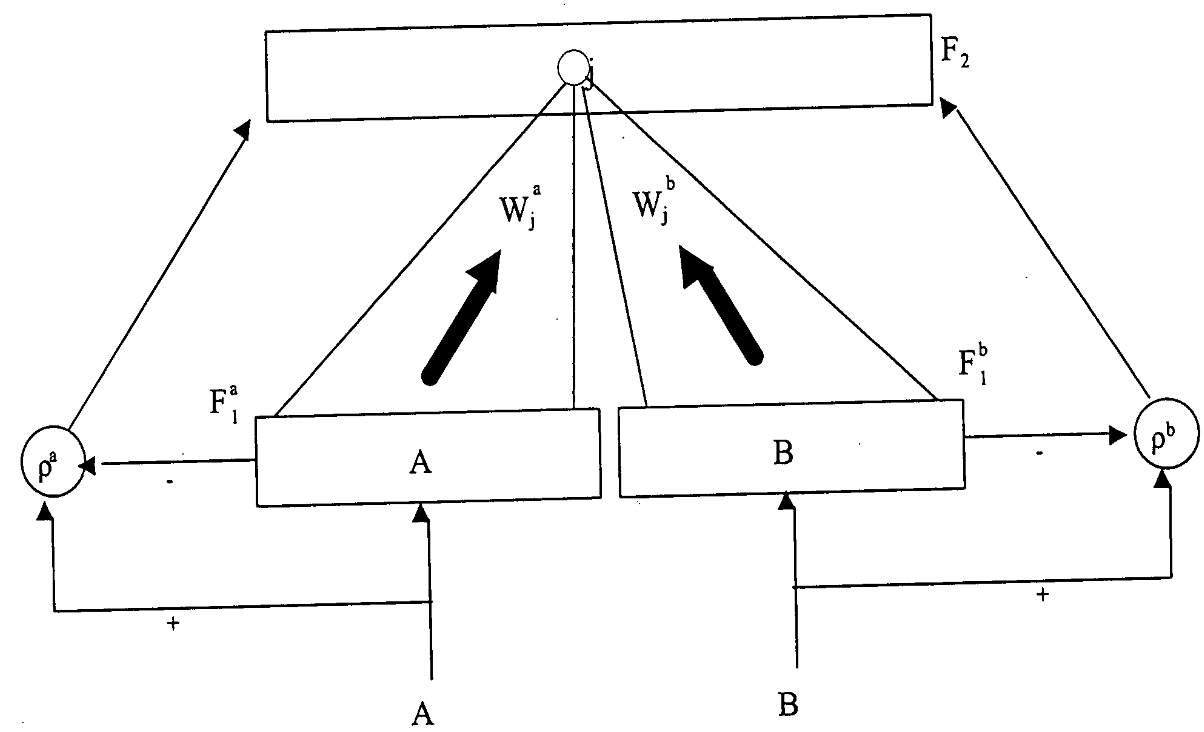


FIG. 2C

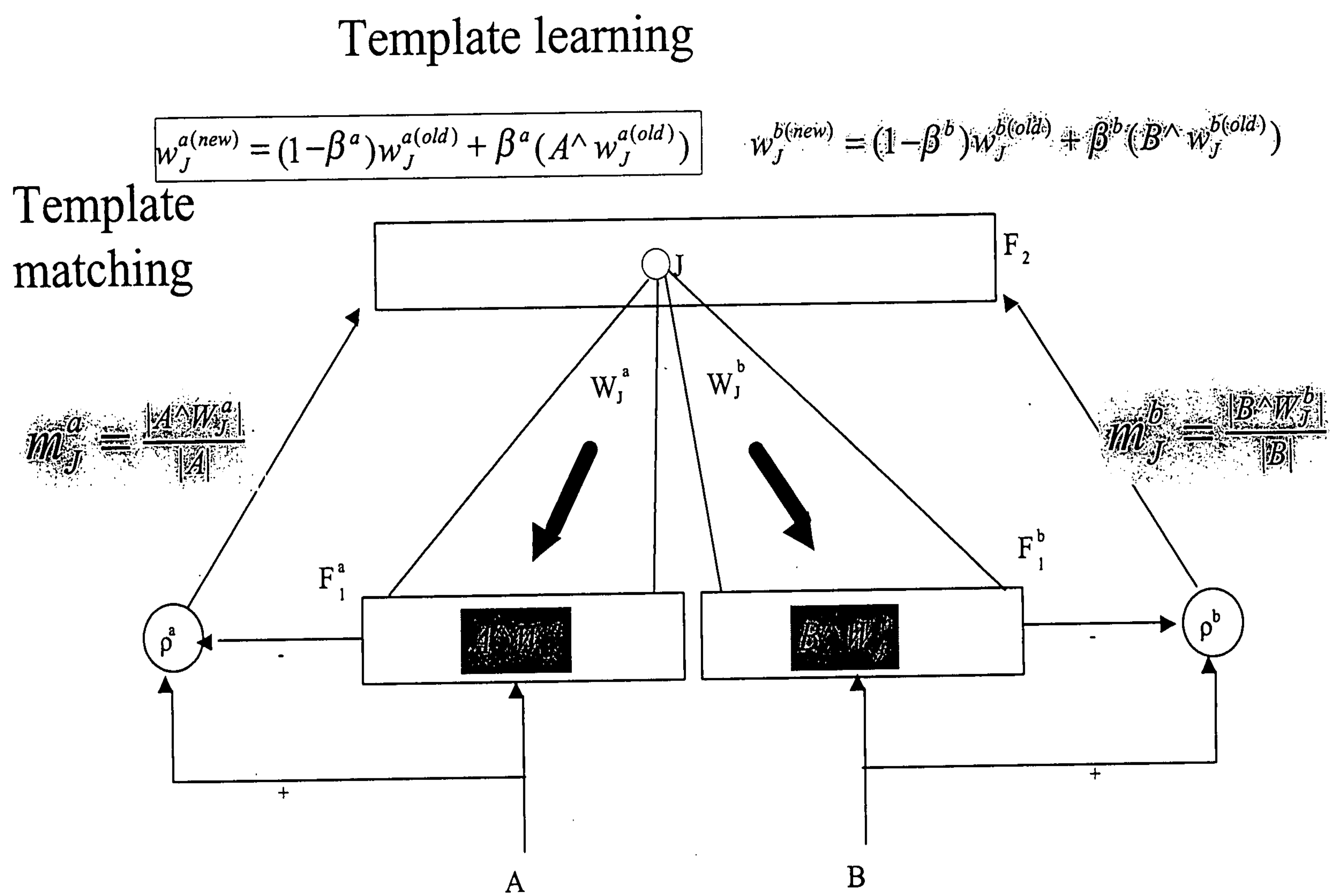

Template learning
$w_J^{a(new)} = (1-\beta^a)w_J^{a(old)} + \beta^a(A \wedge w_J^{a(old)})$
$w_J^{b(new)} = (1-\beta^b)w_J^{b(old)} + \beta^b(B \wedge w_J^{b(old)})$
Template matching
$m_J^a = \frac{|A \wedge W_J^a|}{|A|}$
$m_J^b = \frac{|B \wedge W_J^b|}{|B|}$
F_2
J
W_J^a
W_J^b
F_1^a
F_1^b
ρ^a
ρ^b
-
+
A
B

FIG. 3

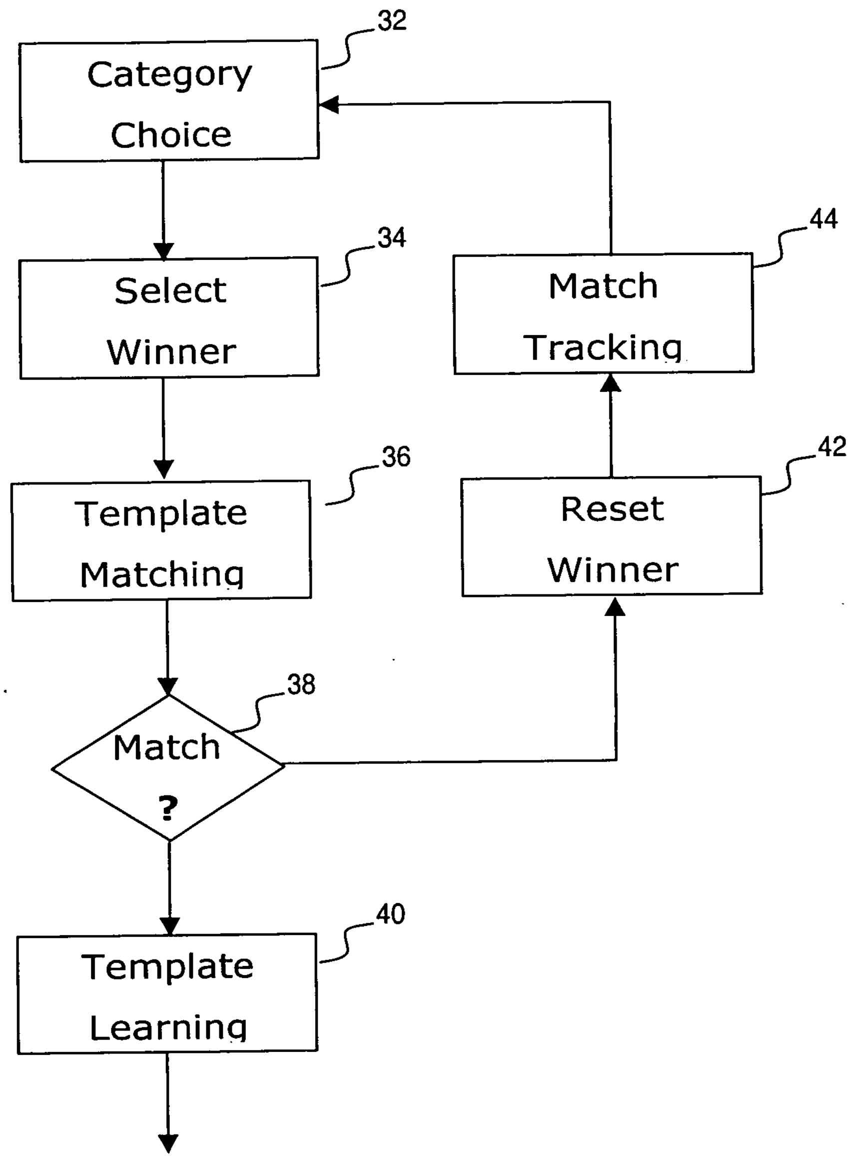
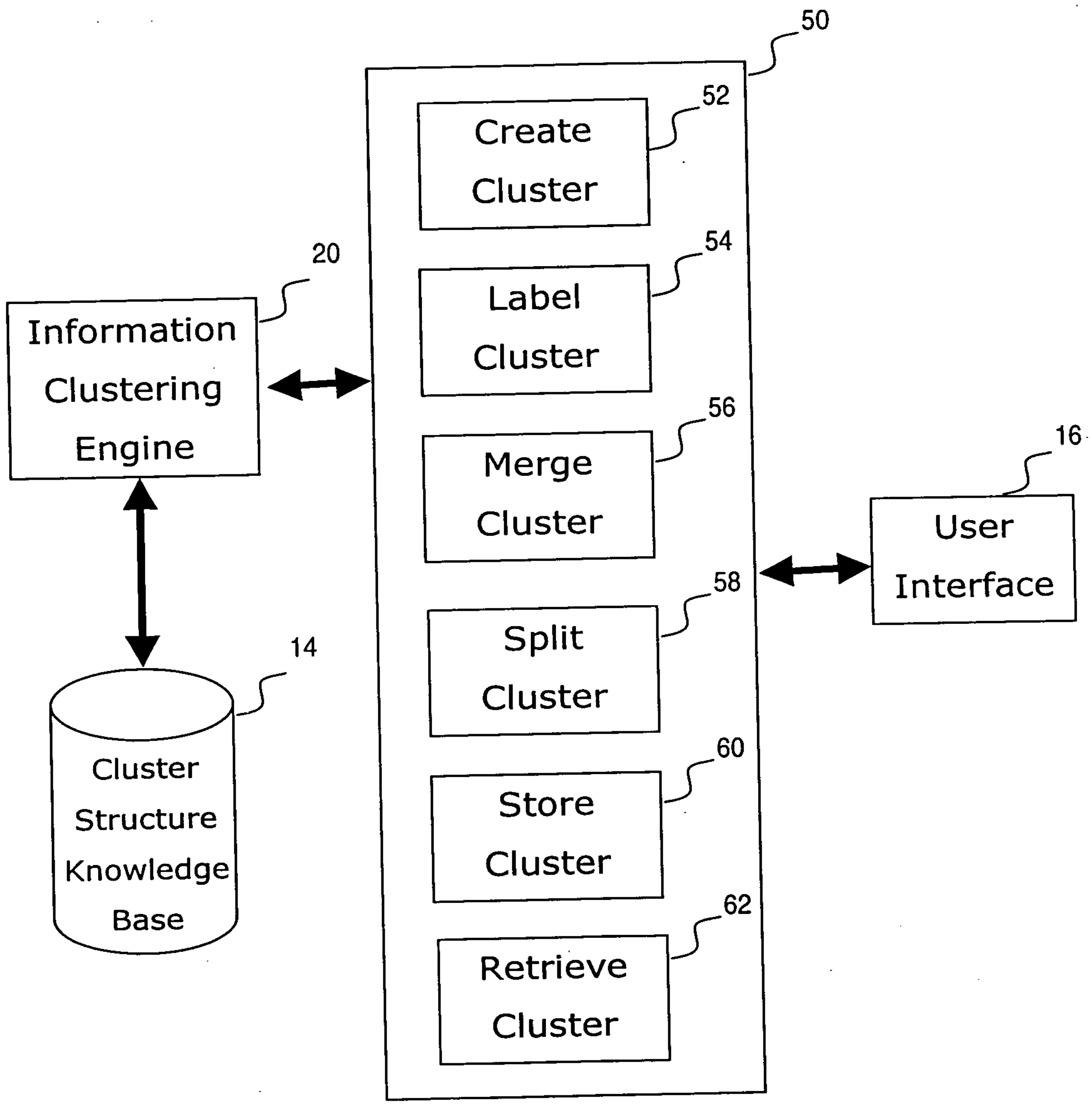

32
Category
Choice
34
Select
Winner
36
Template
Matching
38
Match
?
40
Template
Learning
42
Reset
Winner
44
Match
Tracking

FIG. 4


50
Create Cluster
52
Label Cluster
54
Merge Cluster
56
Split Cluster
58
Store Cluster
60
Retrieve Cluster
62
20
Information Clustering Engine
14
Cluster Structure Knowledge Base
16
User Interface